TOPIC 3

Optimality, convexity, and gradient descent

2 Optimality conditions

Course: Math 535 (http://www.math.wisc.edu/~roch/mmids/) - Mathematical Methods in Data Science (MMiDS)
Author: Sebastien Roch (http://www.math.wisc.edu/~roch/), Department of Mathematics, University of Wisconsin-Madison
Updated: Oct 10, 2020

In this section, we derive optimality conditions for unconstrained continuous optimization problems.

2.1 Multivariate version of Taylor's theorem

We will make use of *Taylor's Theorem*, a powerful generalization of the *Mean Value Theorem* that provides polynomial approximations to a function around a point. We restrict ourselves to the case of a linear approximation with second-order error term, which will suffice for our purposes.

2.1.1 Single-variable case

We begin by reviewing the single-variable case, which we will use to prove the general verison.

Theorem (Taylor): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}$. Suppose f has a continuous derivative on $[a, b]$ and that its second derivative exists on (a, b). Then for any $x \in [a, b]$

$$f(x) = f(a) + (x - a)f'(a) + \frac{1}{2}(x - a)^2 f''(\xi)$$

for some $a < \xi < x$.

Proof idea: The *Mean Value Theorem* implies that there is $a < \xi < x$ such that $f(x) = f(a) + (x-a)f'(\xi)$. One way to think of the proof of that result is the following: we constructed an affine function that agrees with f at a and x, then used *Rolle* to express the coefficient of the linear term using f'. Here we do the same with a polynomial of degree 2. But we now have an extra degree of freedom in choosing this polynomial. Because we are looking for a good approximation close to a, we choose to make the first derivative at a also agree. Applying *Rolle* twice gives the claim.

Proof: Let

$$P(t) = \alpha_0 + \alpha_1(t-a) + \alpha_2(t-a)^2.$$

We choose the α_i's so that $P(a) = f(a)$, $P'(a) = f'(a)$, and $P(x) = f(x)$. The first two lead to the conditions

$$\alpha_0 = f(a), \quad \alpha_1 = f'(a).$$

Let $\phi(t) = f(t) - P(t)$. By construction $\phi(a) = \phi(x) = 0$. By *Rolle*, there is a $\xi' \in (a, x)$ such that $\phi'(\xi') = 0$. Moreover, $\phi'(a) = 0$. Hence we can apply *Rolle* again - this time to ϕ' on $[a, \xi']$. It implies that there is $\xi \in (a, \xi')$ such that $\phi''(\xi) = 0$.

The second derivative of ϕ at ξ is

$$0 = \phi''(\xi) = f''(\xi) - P''(\xi) = f''(\xi) - 2\alpha_2$$

so $\alpha_2 = f''(\xi)/2$. Plugging into P and using $\phi(x) = 0$ gives the claim. □

The third term on the right-hand side of *Taylor's Theorem* is called the remainder. It can be seen as an error term between $f(x)$ and the linear approximation $f(a) + (x-a)f'(a)$. There are other forms (https://en.wikipedia.org/wiki/Taylor%27s_theorem#Explicit_formulas_for_the_remainder) for the remainder. The form we stated here is useful when one has information about the the second derivative. Here is an example.

Example: Consider $f(x) = e^x$. Then $f'(x) = f''(x) = e^x$. Suppose we are interested in approximating f in the interval $[0, 1]$. We take $a = 0$ and $b = 1$ in Taylor's Theorem. The linear term is

$$f(a) + (x-a)f'(a) = 1 + xe^0 = 1 + x.$$

Then for any $x \in [0, 1]$

$$f(x) = 1 + x + \frac{1}{2}x^2 e^{\xi_x}$$

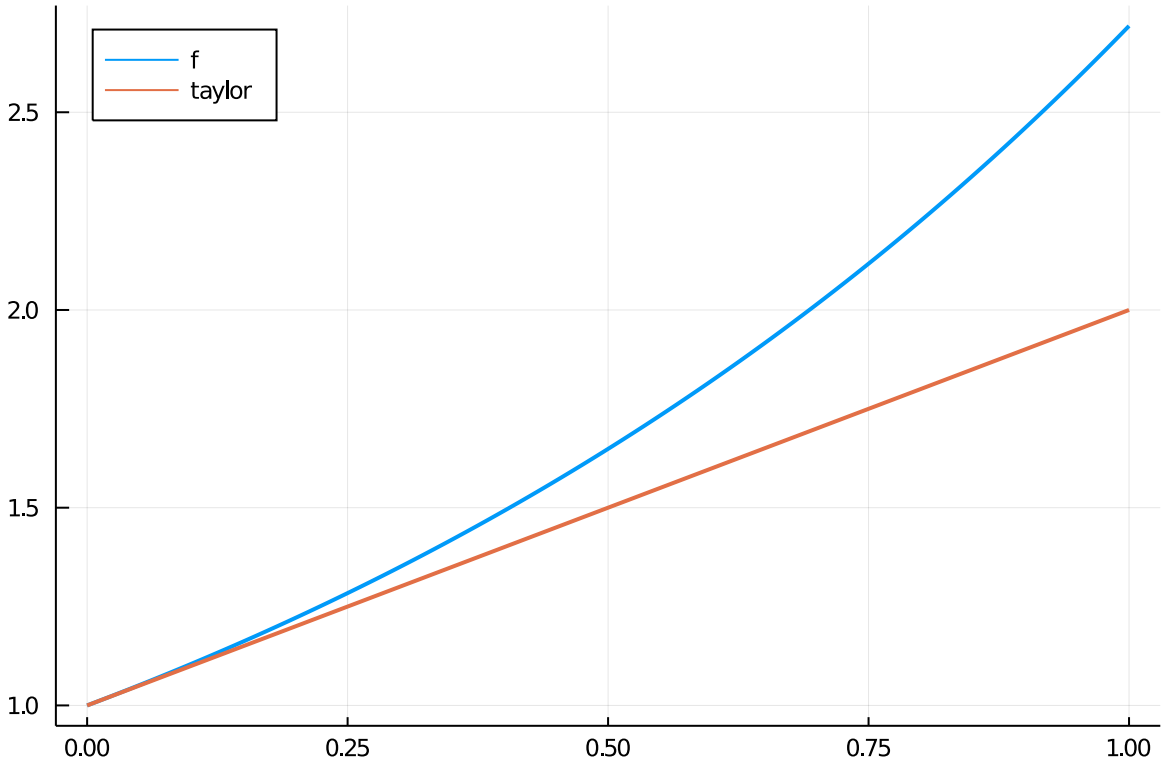
where $\xi_x \in (0, 1)$ depends on x. We get a uniform bound on the error over $[0, 1]$ by replacing ξ_x with its worst possible value over $[0, 1]$

$$|f(x) - (1 + x)| \le \frac{1}{2}x^2 e^{\xi_x} \le \frac{e}{2}x^2.$$

In [1]:
```
# Julia version: 1.5.1
using Plots
```

In [2]:
```
x = LinRange(0,1,100)
y = exp.(x) # function f
taylor = 1 .+ x # linear approximation
err = (exp(1)/2)*x.^2 # our error bound
plot(x, y, lw=2, label="f", legend=:topleft)
plot!(x, taylor, lw=2, label="taylor")
```

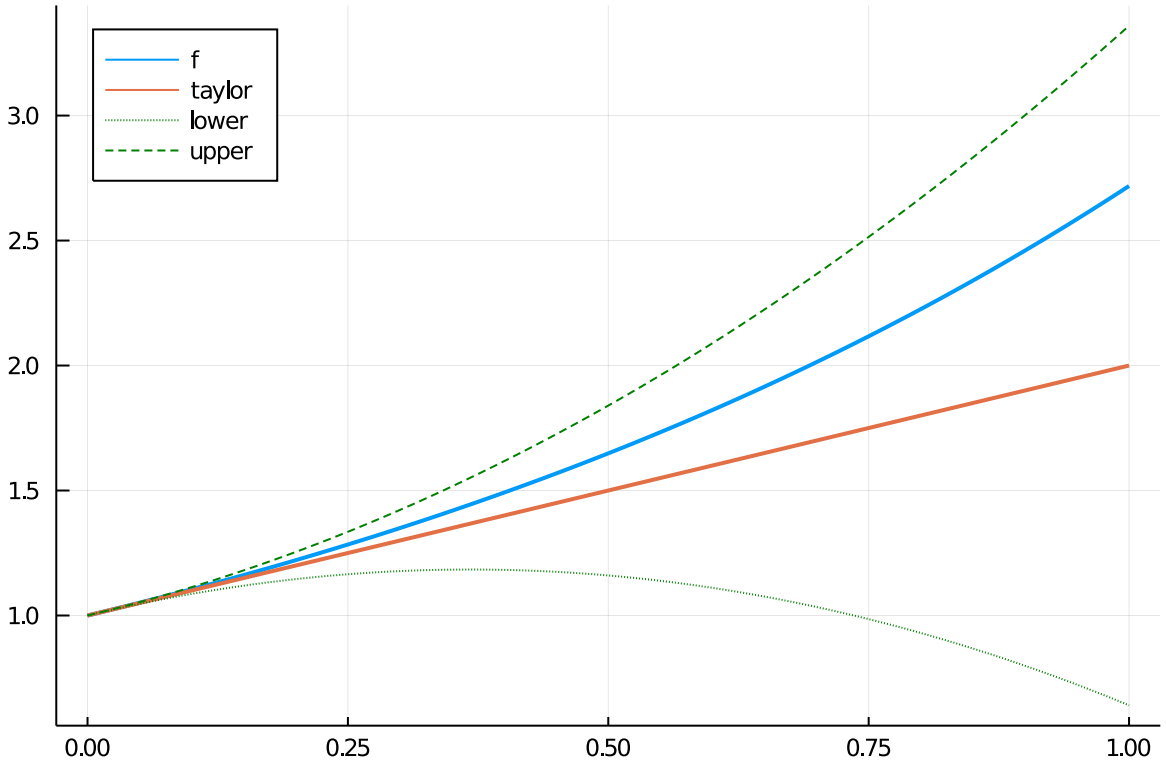
Out[2]:



If we plot the upper and lower bounds, we see that f indeed falls within them.

In [3]:
```
plot!(x, taylor-err, linestyle=:dot, linecolor=:green, label="lower")
plot!(x, taylor+err, linestyle=:dash, linecolor=:green, label="upper")
```

Out[3]:



2.1.2 General case

In the case of several variables, we again restrict ourselves to the second order. We start with an important special case: a *Multivariate Mean Value Theorem*.

Theorem (Multivariate Mean Value): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^d$. Let $\mathbf{x}_0 \in D$ and $\delta > 0$ be such that $B_\delta(\mathbf{x}_0) \subseteq D$. If f is continuously differentiable on $B_\delta(\mathbf{x}_0)$, then for any $\mathbf{x} \in B_\delta(\mathbf{x}_0)$

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0 + \xi\mathbf{p})^T \mathbf{p}$$

for some $\xi \in (0, 1)$, where $\mathbf{p} = \mathbf{x} - \mathbf{x}_0$.

Proof idea: We apply the single-variable result and the *Chain Rule*.

Proof: Let $\phi(t) = f(\boldsymbol{\alpha}(t))$ where $\boldsymbol{\alpha}(t) = \mathbf{x}_0 + t\mathbf{p}$. Observe that $\phi(0) = f(\mathbf{x}_0)$ and $\phi(1) = f(\mathbf{x})$. By the *Chain Rule*,

$$\phi'(t) = \mathbf{J}_f(\boldsymbol{\alpha}(t))\,\mathbf{J}_{\boldsymbol{\alpha}}(t) = \nabla f(\boldsymbol{\alpha}(t))^T\mathbf{p} = \nabla f(\mathbf{x}_0 + t\mathbf{p})^T\mathbf{p}.$$

In particular, ϕ has a continuous first derivative on $[0, 1]$. By the *Mean Value Theorem* in the single-variable case

$$\phi(t) = \phi(0) + t\phi'(\xi)$$

for some $\xi \in (0, t)$. Plugging in the expressions for $\phi(0)$ and $\phi'(\xi)$ and taking $t = 1$ gives the claim. □

We move on to the second-order result. For the more general version, see e.g. Wikipedia (https://en.wikipedia.org/wiki/Taylor's_theorem#Taylor's_theorem_for_multivariate_functions).

Theorem (Multivariate Taylor): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^d$. Let $\mathbf{x}_0 \in D$ and $\delta > 0$ be such that $B_\delta(\mathbf{x}_0) \subseteq D$. If f is twice continuously differentiable on $B_\delta(\mathbf{x}_0)$, then for any $\mathbf{x} \in B_\delta(\mathbf{x}_0)$

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T\mathbf{p} + \frac{1}{2}\mathbf{p}^T\,\mathbf{H}_f(\mathbf{x}_0 + \xi\mathbf{p})\,\mathbf{p}$$

for some $\xi \in (0, 1)$, where $\mathbf{p} = \mathbf{x} - \mathbf{x}_0$.

Proof idea: We apply the single-variable result and the *Chain Rule*.

Proof: Let $\phi(t) = f(\boldsymbol{\alpha}(t))$ where $\boldsymbol{\alpha}(t) = \mathbf{x}_0 + t\mathbf{p}$. Observe that $\phi(0) = f(\mathbf{x}_0)$ and $\phi(1) = f(\mathbf{x})$. As observed in the proof of the *Multivariate Mean Value Theorem*, $\phi'(t) = \nabla f(\boldsymbol{\alpha}(t))^T\mathbf{p}$. By the *Chain Rule*,

$$\phi''(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{i=1}^d \frac{\partial f(\boldsymbol{\alpha}(t))}{\partial x_i}p_i\right] = \sum_{i=1}^d\sum_{j=1}^d \frac{\partial^2 f(\boldsymbol{\alpha}(t))}{\partial x_j \partial x_i}p_j p_i = \mathbf{p}^T\,\mathbf{H}_f(\mathbf{x}_0 + t\mathbf{p})\,\mathbf{p}.$$

In particular, ϕ has continuous first and second derivatives on $[0, 1]$. By *Taylor's Theorem* in the single-variable case

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{1}{2}t^2\phi''(\xi)$$

for some $\xi \in (0, t)$. Plugging in the expressions for $\phi(0)$, $\phi'(0)$ and $\phi''(\xi)$ and taking $t = 1$ gives the claim. □

Example: Consider the function $f(x_1, x_2) = x_1 x_2 + x_1^2 + e^{x_1} \cos x_2$. We apply *Taylor's Theorem* with $\mathbf{x}_0 = (0, 0)$ and $\mathbf{x} = (x_1, x_2)$. The gradient is

$$\nabla f(x_1, x_2) = (x_2 + 2x_1 + e^{x_1} \cos x_2, x_1 - e^{x_1} \sin x_2)^T$$

and the Hessian is

$$\mathbf{H}_f(x_1, x_2) = \begin{pmatrix} 2 + e^{x_1} \cos x_2 & 1 - e^{x_1} \sin x_2 \\ 1 - e^{x_1} \sin x_2 & -e^{x_1} \cos x_2 \end{pmatrix}.$$

So $f(0, 0) = 1$ and $\nabla f(0, 0) = (1, 0)^T$. Thus, by the *Multivariate Taylor's Theorem*, there is $\xi \in (0, 1)$ such that

$$f(x_1, x_2) = 1 + x_1 + \frac{1}{2}[2x_1^2 + 2x_1 x_2 + (x_1^2 - x_2^2)\, e^{\xi x_1} \cos(\xi x_2) - 2x_1 x_2 e^{\xi x_1} \sin(\xi x_2)].$$

◁

2.2 Unconstrained optimization

We will be interested in unconstrained optimization of the form:

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$$

where $f : \mathbb{R}^d \to \mathbb{R}$. In this subsection, we define several notions of solution and derive characterizations.

2.2.1 Definitions

Ideally, we would like to find a global minimizer to the optimization problem above. Recall the definition of a global minimizer.

Definition (Global minimizer): Let $f : \mathbb{R}^d \to \mathbb{R}$. The point $\mathbf{x}^* \in \mathbb{R}^d$ is a global minimizer of f over $\mathbb{R}^d$ if

$$f(\mathbf{x}) \geq f(\mathbf{x}^*), \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

◁

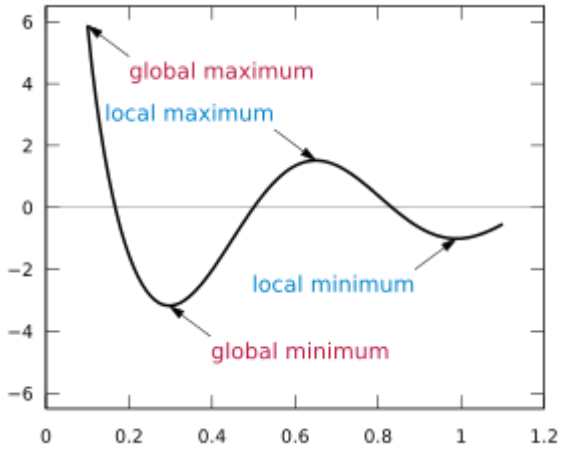
We have observed before that, in general, finding a global minimizer and certifying that one has been found can be difficult unless some special structure is present. Therefore weaker notions of solution are needed. Recall the notion of a local minimizer.

Definition (Local minimizer): Let $f : \mathbb{R}^d \to \mathbb{R}$. The point $\mathbf{x}^* \in \mathbb{R}^d$ is a local minimizer of f over $\mathbb{R}^d$ if there is $\delta > 0$ such that

$$f(\mathbf{x}) \geq f(\mathbf{x}^*), \quad \forall \mathbf{x} \in B_\delta(\mathbf{x}^*) \setminus \{\mathbf{x}^*\}.$$

If the inequality is strict, we say that $\mathbf{x}^*$ is a strict local minimizer. ◁

In words, $\mathbf{x}^*$ is a local minimizer if there is open ball around $\mathbf{x}^*$ where it attains the minimum value. The difference between global and local minimizers is illustrated in the next figure.



(Source (https://commons.wikimedia.org/wiki/File:Extrema_example_original.svg))

2.2.2 Necessary conditions

Local minimizers can be characterized in terms of the gradient and Hessian of the function.

We first generalize the *Descent Direction Lemma* to the multivariate case. We first need to define what a descent direction is.

Definition (Descent Direction): Let $f : \mathbb{R}^d \to \mathbb{R}$. A vector $\mathbf{v}$ is a descent direction for f at $\mathbf{x}_0$ if there is $\alpha^* > 0$ such that

$$f(\mathbf{x}_0 + \alpha \mathbf{v}) < f(\mathbf{x}_0), \quad \forall \alpha \in (0, \alpha^*).$$

◁

In the continuously differentiable case, the directional derivative gives a criterion for descent directions.

Lemma (Descent Direction and Directional Derivative): Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable at $\mathbf{x}_0$. A vector $\mathbf{v}$ is a descent direction for f at $\mathbf{x}_0$ if

$$\frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{v}} = \nabla f(\mathbf{x}_0)^T \mathbf{v} < 0$$

that is, if the directional derivative of f at $\mathbf{x}_0$ in the direction $\mathbf{v}$ is negative.

Proof idea: We use the *Multivariate Mean Value Theorem* to show that f takes smaller values in direction $\mathbf{v}$.

Proof: Suppose there is $\mathbf{v} \in \mathbb{R}^d$ such that $\nabla f(\mathbf{x}_0)^T \mathbf{v} = -\eta < 0$. For $\alpha > 0$, the *Multivariate Mean Value Theorem* implies that there is $\xi_\alpha \in (0, 1)$ such that

$$f(\mathbf{x}_0 + \alpha \mathbf{v}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0 + \xi_\alpha \alpha \mathbf{v})^T (\alpha \mathbf{v}) = f(\mathbf{x}_0) + \alpha \nabla f(\mathbf{x}_0 + \xi_\alpha \alpha \mathbf{v})^T \mathbf{v}.$$

We cannot immediately apply our condition on $\mathbf{v}$ as the gradient in the previous equation is taken at $\mathbf{x}_0 + \xi_\alpha \alpha \mathbf{v}$, not $\mathbf{x}_0$. But the gradient is continuous, so there is $\delta > 0$ such that

$$\left| \nabla f(\mathbf{x})^T \mathbf{v} - \nabla f(\mathbf{x}_0)^T \mathbf{v} \right| < \eta/2$$

for all $\mathbf{x} \in B_\delta(\mathbf{x}_0)$. Hence there is $\alpha^* > 0$ small enough such that

$$\nabla f(\mathbf{x}_0 + \xi_\alpha \alpha \mathbf{v})^T \mathbf{v} < -\eta/2 < 0, \quad \forall \alpha \in (0, \alpha^*).$$

That implies

$$f(\mathbf{x}_0 + \alpha \mathbf{v}) < f(\mathbf{x}_0) - \alpha \eta/2 < f(\mathbf{x}_0), \quad \forall \alpha \in (0, \alpha^*)$$

and proves the claim. □

Lemma (Descent Direction: Multivariate Version): Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable at $\mathbf{x}_0$ and assume that $\nabla f(\mathbf{x}_0) \neq 0$. Then f has a descent direction at $\mathbf{x}_0$.

Proof: Take $\mathbf{v} = -\nabla f(\mathbf{x}_0)$. Then $\nabla f(\mathbf{x}_0)^T \mathbf{v} = -\|\nabla f(\mathbf{x}_0)\|^2 < 0$ since $\nabla f(\mathbf{x}_0) \neq 0$. □

This leads to the following fundamental result.

Theorem (First-Order Necessary Condition): Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable on $\mathbb{R}^d$. If $\mathbf{x}_0$ is a local minimizer, then $\nabla f(\mathbf{x}_0) = 0$.

Proof idea: In a descent direction, f decreases hence there cannot be one at a local minimizer.

Proof: We argue by contradiction. Suppose that $\nabla f(\mathbf{x}_0) \neq 0$. By the *Descent Direction Lemma*, there is a descent direction $\mathbf{v} \in \mathbb{R}^d$ at $\mathbf{x}_0$. That implies

$$f(\mathbf{x}_0 + \alpha \mathbf{v}) < f(\mathbf{x}_0), \quad \forall \alpha \in (0, \alpha^*)$$

for some $\alpha^* > 0$. So every open ball around $\mathbf{x}_0$ has a point achieving a smaller value than $f(\mathbf{x}_0)$. Thus $\mathbf{x}_0$ is not a local minimizer, a contradiction. So it must be that $\nabla f(\mathbf{x}_0) = 0$. □

When f is twice continuously differentiable, we also get a condition on the Hessian.

Theorem (Second-Order Necessary Condition): Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable on $\mathbb{R}^d$. If $\mathbf{x}_0$ is a local minimizer, then $\mathbf{H}_f(\mathbf{x}_0)$ is positive semidefinite.

Proof idea: The proof goes along the same lines as the argument leading to the *First-Order Necessary Condition*, but we use the *Multivariate Taylor's Theorem* to the second order instead.

Proof: We argue by contradiction. Suppose that $\mathbf{H}_f(\mathbf{x}_0)$ is not positive semidefinite. From the *Symmetry of the Hessian Theorem* and the *Spectral Theorem*, $\mathbf{H}_f(\mathbf{x}_0)$ has a spectral decomposition. From the *Characterization of Positive Semidefiniteness*, however, it follows that $\mathbf{H}_f(\mathbf{x}_0)$ must have at least one negative eigenvalue $-\eta < 0$. Let $\mathbf{v}$ be a corresponding eigenvector.

We have that $\langle \mathbf{v}, \mathbf{H}_f(\mathbf{x}_0)\, \mathbf{v} \rangle = -\eta < 0$. For $\alpha > 0$, the *Multivariate Taylor's Theorem* implies that there is $\xi_\alpha \in (0, 1)$ such that

$$\begin{aligned} f(\mathbf{x}_0 + \alpha \mathbf{v}) &= f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T (\alpha \mathbf{v}) + (\alpha \mathbf{v})^T \mathbf{H}_f(\mathbf{x}_0 + \xi_\alpha \alpha \mathbf{v})(\alpha \mathbf{v}) \\ &= f(\mathbf{x}_0) + \alpha^2 \mathbf{v}^T \mathbf{H}_f(\mathbf{x}_0 + \xi_\alpha \alpha \mathbf{v})\, \mathbf{v} \end{aligned}$$

where we used $\nabla f(\mathbf{x}_0) = 0$ by the *First-Order Necessary Condition*.

Since the Hessian is continuous, there is $\delta > 0$ such that

$$\left| \mathbf{v}^T \mathbf{H}_f(\mathbf{x})\, \mathbf{v} - \mathbf{v}^T \mathbf{H}_f(\mathbf{x}_0)\, \mathbf{v} \right| < \eta/2$$

for all $\mathbf{x} \in B_\delta(\mathbf{x}_0)$. So taking α small enough gives

$$\mathbf{v}^T \mathbf{H}_f(\mathbf{x}_0 + \xi_\alpha \alpha \mathbf{v})\, \mathbf{v} < -\eta/2 < 0.$$

That implies

$$f(\mathbf{x}_0 + \alpha \mathbf{v}) < f(\mathbf{x}_0) - \alpha^2 \eta/2 < f(\mathbf{x}_0).$$

Since this holds for all sufficiently small α, every open ball around $\mathbf{x}_0$ has a point achieving a lower value than $f(\mathbf{x}_0)$. Thus $\mathbf{x}_0$ is not a local minimizer, a contradiction. So it must be that $\mathbf{H}_f(\mathbf{x}_0) \succeq 0$. □

2.2.3 Sufficient conditions

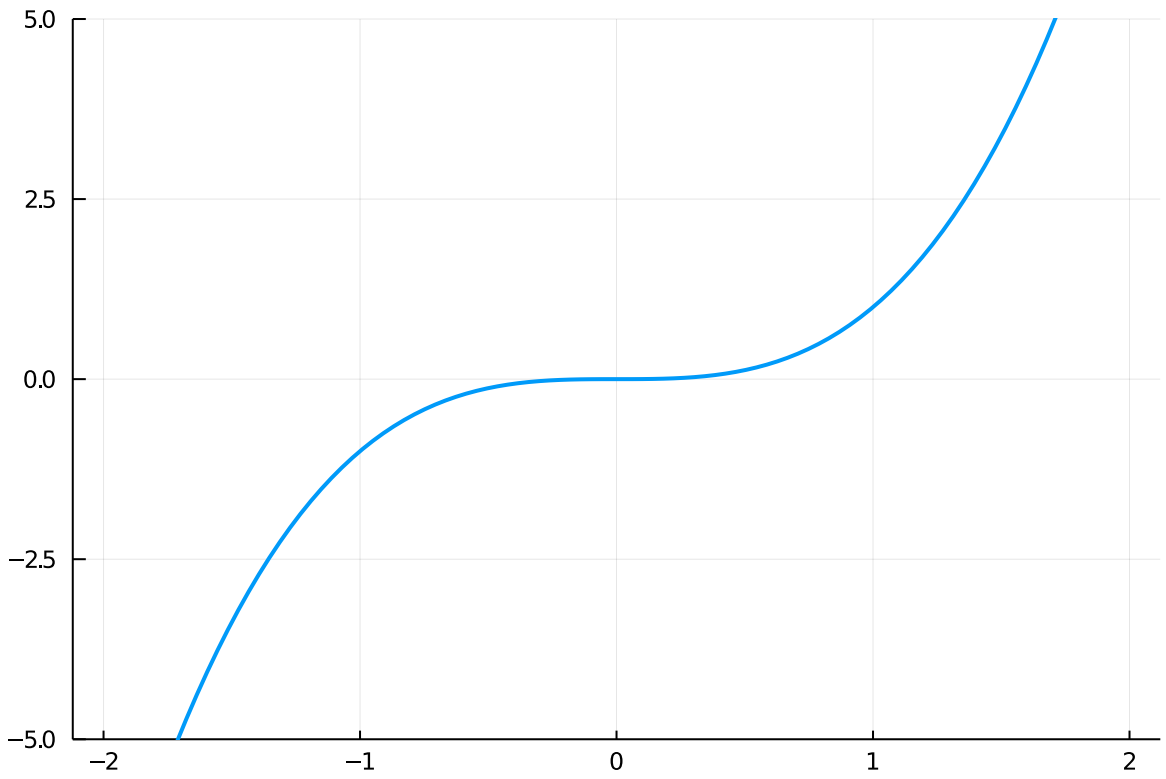
The necessary conditions in the previous subsection are not in general sufficient, as the following example shows.

Definition (Stationary Point): Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable on $\mathbb{R}^d$. If $\nabla f(\mathbf{x}_0) = 0$, we say that $\mathbf{x}_0$ is a stationary point of f.◁

Example: Let $f(x) = x^3$. Then $f'(x) = 3x^2$ and $f''(x) = 6x$ so that $f'(0) = 0$ and $f''(0) \geq 0$. Hence $x = 0$ is a stationary point. But $x = 0$ is not a local minimizer. Indeed $f(0) = 0$ but, for any $\delta > 0$, $f(-\delta) < 0$.

In [4]:
```
f(x) = x^3
x = LinRange(-2,2, 100)
y = f.(x)
plot(x, y, lw=2, legend=false, ylim = (-5,5))
```

Out[4]:



◁

We give sufficient conditions for a local minimizer. We will need the following lemma, whose proof relies on the next exercises.

Exercise: Prove the following claim, which is known as the *Subspace Intersection Lemma*. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be linear subspaces of $\mathbb{R}^d$ and let

$$\mathcal{S}_1 + \mathcal{S}_2 = \{\mathbf{x}_1 + \mathbf{x}_2 \ : \ \forall \mathbf{x}_1 \in \mathcal{S}_1, \mathbf{x}_2 \in \mathcal{S}_2\}.$$

Then it holds that

$$\dim(\mathcal{S}_1 + \mathcal{S}_2) = \dim(\mathcal{S}_1) + \dim(\mathcal{S}_2) - \dim(\mathcal{S}_1 \cap \mathcal{S}_2).$$

[Hint: Consider a basis of $\mathcal{S}_1 \cap \mathcal{S}_2$ and complete into bases of $\mathcal{S}_1$ and $\mathcal{S}_2$. Show that the reuslting list of vectors is linear independent.] ⊲

Exercise: Show that, for any linear subspaces $\mathcal{S}_1, \dots, \mathcal{S}_m$ of $\mathcal{V} = \mathbb{R}^d$, it holds that

$$\dim\left(\bigcap_{k=1}^{m} \mathcal{S}_k\right) \geq \sum_{k=1}^{m} \dim\left(\mathcal{S}_k\right) - (m-1)\dim(\mathcal{V}).$$

[Hint: Use the Subspace Intersection Lemma and induction.] ⊲

For a symmetric matrix $C \in \mathbb{R}^{d\times d}$, we let $\lambda_j(C)$, $j = 1, \dots, d$, be the eigenvalues of C in non-increasing order with corresponding orthonormal eigenvectors $\mathbf{v}_j$, $j = 1, \dots, d$. Define the subspaces

$$\mathcal{V}_k(C) = \operatorname{span}(\mathbf{v}_1, \dots, \mathbf{v}_k) \quad \text{and} \quad \mathcal{W}_{d-k+1}(C) = \operatorname{span}(\mathbf{v}_k, \dots, \mathbf{v}_d).$$

The following lemma is one version of what is known as Weyl's Inequality.

Lemma (Weyl's Inequality): Let $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times d}$ be symmetric matrices. Then, for all $j = 1, \dots, d$,

$$\max_{j\in[d]} \left|\lambda_j(B) - \lambda_j(A)\right| \leq \|B - A\|_2$$

where $\|C\|_2$ is the induced 2-norm of C.

Proof idea: We use the extremal characterization of the eigenvalues together with a dimension argument.

Proof: Let $H = B - A$. We prove only the upper bound. The other direction follows from interchanging the roles of A and B. Because

$$\dim(\mathcal{V}_j(B)) + \dim(\mathcal{W}_j(A)) + \dim(\mathcal{W}_1(H)) = j + (d - j + 1) + d = 2d + 1$$

it follows from the exercise above that

$$\dim\left(\mathcal{V}_j(B) \cap \mathcal{W}_j(A) \cap \mathcal{W}_1(H)\right) \geq (2d+1) - 2d = 1.$$

Hence the $\mathcal{V}_j(B) \cap \mathcal{W}_j(A) \cap \mathcal{W}_1(H)$ is non-empty. Let $\mathbf{v}$ be a unit vector in that intersection.

By *Courant-Fischer*,

$$\lambda_j(B) \leq \langle \mathbf{v}, (A+H)\mathbf{v} \rangle = \langle \mathbf{v}, A\mathbf{v} \rangle + \langle \mathbf{v}, H\mathbf{v} \rangle \leq \lambda_j(A) + \langle \mathbf{v}, H\mathbf{v} \rangle.$$

Moreover, by *Cauchy-Schwarz*, since $\|\mathbf{v}\| = 1$

$$\langle \mathbf{v}, H\mathbf{v} \rangle \leq \|\mathbf{v}\| \|H\mathbf{v}\| \leq \|H\|_2$$

which proves the claim. □

Theorem (Second-Order Sufficient Condition): Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable on $\mathbb{R}^d$. If $\nabla f(\mathbf{x}_0) = \mathbf{0}$ and $\mathbf{H}_f(\mathbf{x}_0)$ is positive definite, then $\mathbf{x}_0$ is a strict local minimizer.

Proof idea: We use the *Multivariate Taylor's Theorem* again. This time we use the positive definiteness of the Hessian to bound the value of the funciton from below. We use *Weyl's Inequality* to show that the eigenvalues of the Hessian are continuous, which implies that the Hessian remains positive definite in an open ball around $\mathbf{x}_0$.

Proof: We first claim that there is $\rho > 0$ such that $\mathbf{H}_f(\mathbf{x})$ is positive definite for all $\mathbf{x} \in B_\rho(\mathbf{x}_0)$. That is the case when $\mathbf{x} = \mathbf{x}_0$ and let $\mu_1 > 0$ be the smallest eigenvalue $\mathbf{H}_f(\mathbf{x}_0)$. We use *Weyl's Inequality* to bound the eigenvalues of the Hessian from below around $\mathbf{x}_0$. For any vector $\mathbf{v} \in \mathbb{R}^d$, we have for all $j = 1, \ldots, d$

$$\lambda_j(\mathbf{H}_f(\mathbf{x}_0 + \mathbf{v})) \geq \lambda_j(\mathbf{H}_f(\mathbf{x}_0)) - \|\mathbf{H}_f(\mathbf{x}_0 + \mathbf{v}) - \mathbf{H}_f(\mathbf{x}_0)\|_2$$

where we used the notation introduced above. We bound $\lambda_j(\mathbf{H}_f(\mathbf{x}_0)) \geq \lambda_1(\mathbf{H}_f(\mathbf{x}_0)) = \mu_1$ and

$$\|\mathbf{H}_f(\mathbf{x}_0 + \mathbf{v}) - \mathbf{H}_f(\mathbf{x}_0)\|_2 \leq \|\mathbf{H}_f(\mathbf{x}_0 + \mathbf{v}) - \mathbf{H}_f(\mathbf{x}_0)\|_F.$$

The Frobenius norm above is continuous in $\mathbf{v}$ as a composition of continuous functions. Moreover, we of course have at $\mathbf{v} = \mathbf{0}$ that this Frobenius is 0. Hence, by definition of continuity, there is $\rho > 0$ such that for all $\mathbf{v} \in B_\rho(\mathbf{0})$ we have $\|\mathbf{H}_f(\mathbf{x}_0 + \mathbf{v}) - \mathbf{H}_f(\mathbf{x}_0)\|_F < \mu_1/2$. Plugging back above finally gives for such $\mathbf{v}$'s and all $j = 1, \ldots, d$

$$\lambda_j(\mathbf{H}_f(\mathbf{x}_0 + \mathbf{v})) > \mu_1/2 > 0.$$

In particular, $\mathbf{H}_f(\mathbf{x}_0 + \mathbf{v})$ is positive definite and $\langle \mathbf{u}, \mathbf{H}_f(\mathbf{x}_0 + \mathbf{v})\, \mathbf{u} \rangle > \frac{\mu_1}{2}\|\mathbf{u}\|^2$ for any $\mathbf{u}$ by *Courant-Fischer*.

By the *Multivariate Taylor's Theorem*, $\forall \mathbf{v} \in B_\rho(\mathbf{0}) \setminus \{\mathbf{0}\}$ there is $\xi \in (0, 1)$

$$\begin{aligned} f(\mathbf{x}_0 + \mathbf{v}) &= f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T \mathbf{v} + \mathbf{v}\, \mathbf{H}_f(\mathbf{x}_0 + \xi\mathbf{v})\, \mathbf{v} \\ &> f(\mathbf{x}_0) + \frac{\mu_1}{2}\|\mathbf{v}\|^2 \\ &> f(\mathbf{x}_0) \end{aligned}$$

where we used that $\|\xi\mathbf{v}\| = \xi\|\mathbf{v}\| \leq \rho$. Therefore $\mathbf{x}_0$ is a strict local minimizer. □